Activation Functions Explained

- Activation functions are used to activate the neurons in the model.
- Activation functions are used to learn the non-linearity in the data.

- Following properties are considered while selecting the activation functions:
    1. Non-Linear
    2. Differentiable
    3. Computationally Inexpensive
    4. Zero-Centered (**Normalized or mean=0, balanced in +ve and -ve**)
    5. Non-Saturating (should not squash in specific range)

Reference:

https://www.youtube.com/watch?v=7LcUkgzx3AY (part1)

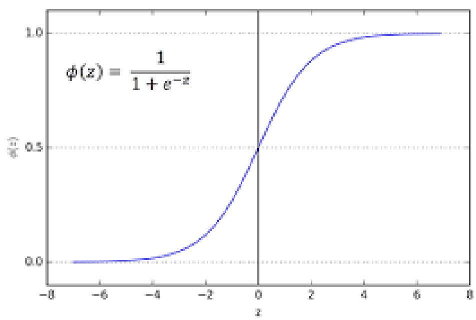
https://www.youtube.com/watch?v=2OwWs7Hzr9g (part2)

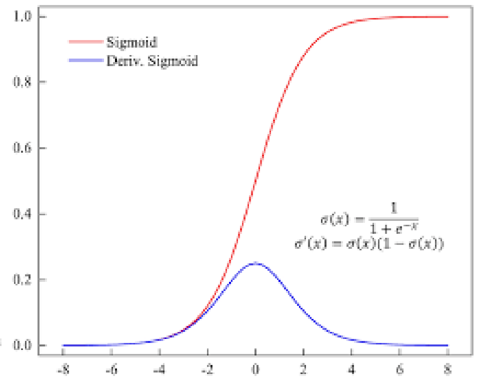
1. NOTE: For relu, leaky relu, PReLU , eLU, SeLU watch part2 of video

Types of Activation Function:

2. Sigmoid
3. Tanh
4. ReLU
5. Leaky ReLU
6. PReLU (Parametric relu)
7. eLU
8. SeLU
9. Softmax

1. Sigmoid (aka Logistic Activation Function):





- The Sigmoid Function curve looks like an S-shape.
- This function takes any real value as input and outputs values in the range of 0 to 1.
- The larger the input (more positive), the closer the output value will be to 1.0, whereas the smaller the input (more negative), the closer the output will be to 0.0
- Mathematically sigmoid function can be written as,

$$\textit{Sigmoid / Logistic}$$

$$f(x) = \frac{1}{1+e^{-x}}$$

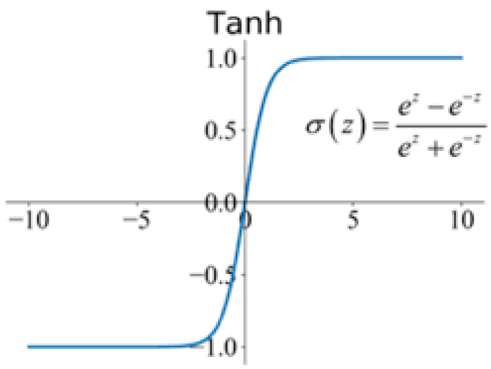
Pros:

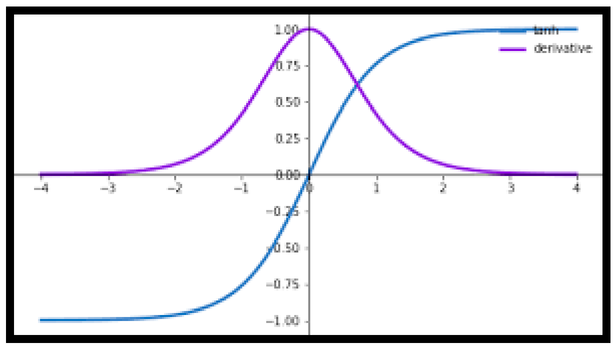
- This sigmoid function is differentiable.
- As it ranges in (0, 1) so it can be used in binary classification use case.
- It adds non-linearity while learning the data pattern.

Cons:

- Squashing function, saturating function (in certain range), which leads to Vanishing Gradient Problem.
- Computationally expensive due to exponential term in it.
- Not zero centered.
- Leads to "**Vanishing gradient**" problem

2. Tanh (aka Hyperbolic Tangent):





Tanh

$$f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$

- The tanh function became preferred over the sigmoid function as it gave better performance for multi-layer neural networks. But it did not solve the vanishing gradient problem that sigmoid suffered, which was tackled more effectively with the introduction of ReLU activations.
- Tanh function is very similar to the sigmoid/logistic activation function, and even has the same S-shape with the difference in output range of -1 to 1. In Tanh, the larger the input (more positive), the closer the output value will be to 1.0, whereas the smaller the input (more negative), the closer the output will be to -1.0.

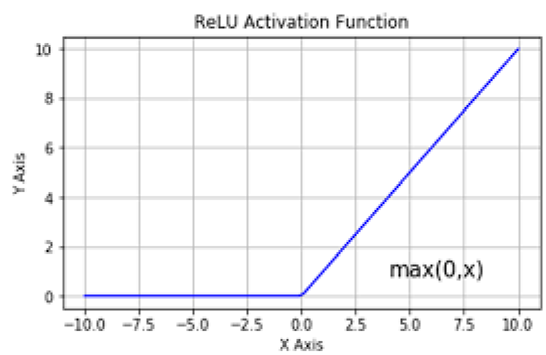
Pros:

- Non-Linear
- Differentiable
- Zero-Centered (Having Normalized values, and thus faster training compared to sigmoid)

Cons:

- Saturating/squashing function.
- Computationally expensive due to exponential term.
- Leads to "**Vanishing gradient**" problem

3. Relu (aka Rectified Linear unit):



- Relu is a Non-linear function due to max component, even if it seems to be linear.
- Mathematically expressed as *f(x) = max (0, x)*

Pros:

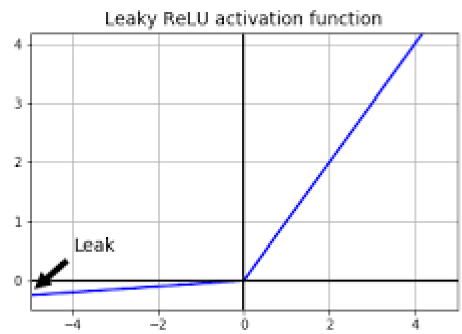
- Non-Linear
- Not saturated in positive region
- Computationally Inexpensive

Cons:

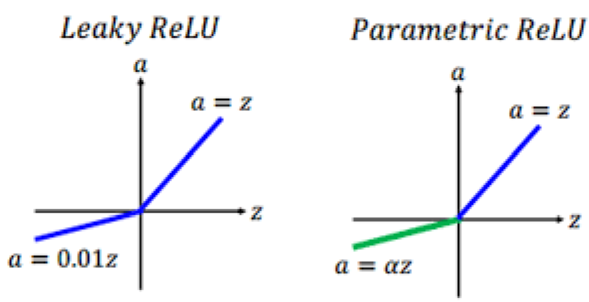
- Saturating/squashing function for negative inputs.
- Not differentiable at 0, thus, if x >= 0, f(x) = 1, else, f(x) = 0

- Not zero centered. (to avoid this, we use batch normalization)
- Leads to "**Dying Relu**" problem

4. Leaky Relu :



5. PReLU :



6. eLU :

7. SeLU :

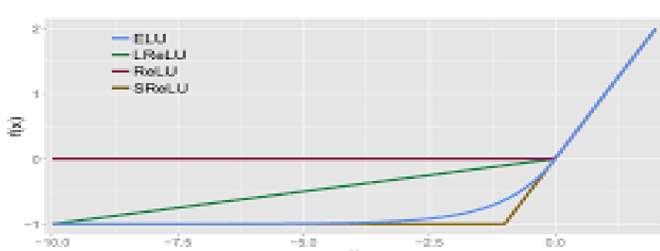
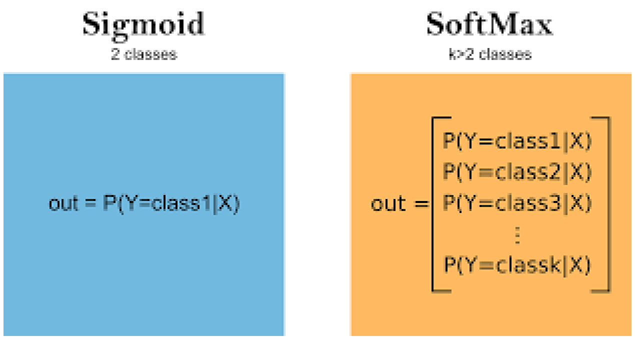


Figure 1: The rectified linear unit (ReLU), the leaky ReLU (LReLU, $\alpha = 0.1$), the shifted ReLUs (SReLUs), and the exponential linear unit (ELU, $\alpha = 1.0$).

8. Sigmoid:



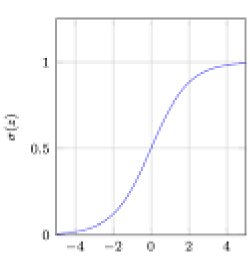
$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

Example :

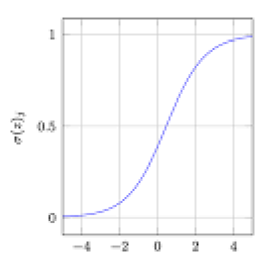
$$2.33 \longrightarrow P\,(\text{Class } 1) = \frac{\exp(2.33)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.83827314$$

$$-1.46 \longrightarrow P\,(\text{Class } 2) = \frac{\exp(-1.46)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.01894129$$

$$0.56 \longrightarrow P\,(\text{Class } 3) = \frac{\exp(0.56)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.14278557$$



(a) Sigmoid activation function.



(b) Softmax activation function.